JSC370 2026 Final Project Report

Yi Fan (Eric) Wang

2026-04-26

---

1. Introduction

Climate change is one of the major global challenges and is largely driven by carbon dioxide emissions from human activities. These emissions are not evenly distributed across countries: countries at different stages of economic development rely on different energy sources, follow different urbanization trends, and face different land-use pressures, all of which leave distinct imprints on their per-capita emissions. Understanding what drives these country-level differences is important for designing effective climate and energy policies.

This project examines the following research question:

How have carbon dioxide emissions per capita evolved across countries from 1990 to 2023, and which economic, energy, and land-use factors explain and predict these differences?

Based on the midterm analysis, this report hypothesizes that GDP per capita, urban population share, and renewable energy share are the dominant cross-country drivers of carbon dioxide emissions per capita, while fossil-fuel electricity share, and the calendar year play a smaller role.

The analysis uses data from 1990 to 2023 from the World Bank Open Data API, which provides a comprehensive set of development indicators covering economic, energy, and land-use factors. The dataset includes carbon dioxide emissions per capita, GDP per capita, renewable energy share, fossil-fuel electricity share, forest area as a percentage of land area, and urban population share, along with the year of observation as a temporal indicator.

This report extends the midterm analysis with more flexible methods that can capture more complex nonlinearities and interactions, and reports both predictive performance and variable importance to identify the dominant drivers of cross-country differences, as the midterm analysis found statistically significant evidence of non-linear relationships between carbon dioxide emissions per capita and its predictors.

2. Methods

2.1 Data Source and Acquisition

Data were obtained from the World Bank Open Data API for 1990–2023. The indicators selected are carbon dioxide emissions per capita, GDP per capita, renewable energy share, fossil-fuel electricity share, forest area as a percentage of land area, and urban population share.

2.2 Data Cleaning and Wrangling

The data were indexed by ISO3 country code and year to form a country–year panel. Regional aggregates were removed. The dataset was reshaped to wide format, variable types and ranges were checked, and missing values were recorded as `NaN`. Missing observations were kept during EDA but dropped before regression and modelling.

2.3 Exploratory Data Analysis

EDA used summary statistics, line plots, histograms, and scatter plots to examine temporal trends, distributional properties, and associations between emissions per capita and the explanatory variables.

2.4 Feature Engineering

Log transformations were applied to right-skewed variables (carbon dioxide emissions per capita and GDP per capita) to reduce the influence of extreme values. Percentage variables were kept on their original scale. The year was centered at 1990 to make the intercept and model coefficients more interpretable.

2.5 Regression Analysis

Before modelling, rows with missing values were removed so that all models used the same complete dataset.

OLS and GAM models from the midterm were used to capture linear and nonlinear relationships. The nonlinear patterns identified by the GAM motivated the use of more flexible machine learning models in capturing more complex nonlinear patterns in the next step.

2.6 Modeling

Because the GAM showed statistically significant nonlinear relationships, two tree-based models were added. A pruned Decision Tree Regressor was used as an interpretable baseline. Then a XGBoost Regressor was used as a flexible gradient-boosted ensemble that can capture more complex nonlinearities and interactions.

2.6.1 Train / Validation / Test Split

The dataset is a country–year panel, so observations were split at the country level (grouped by `countryiso3code`) to prevent leakage of country-specific effects across sets. The split was approximately 70% / 15% / 15% by country, with the same fixed random seed across models for reproducibility and fair comparison.

Since entire countries were held out, test performance reflects generalization to unseen countries using their full 1990–2023 data. This setup does not evaluate time-based forecasting.

2.6.2 Decision Tree Model

A Decision Tree Regressor was fit as the interpretable baseline. The cost-complexity pruning parameter (α) was tuned on the validation set, and the final pruned tree was refit on the combined training and validation sets and evaluated once on the test set.

2.6.3 XGBoost Model

XGBoost was tuned by randomized search over tree depth, learning rate, regularization, row/column subsampling, and minimum child weight, using 5-fold country-grouped cross-validation on the training set. The number of boosting rounds was then selected by cross-validated early stopping on the training set. The final tuned model was refit on the combined training and validation data and evaluated on the test set. An untuned XGBoost baseline (200 trees, max depth 5, learning rate 0.1) was also trained for comparison.

2.6.4 Model Evaluation and Comparison

Models were compared on the same test set using regression model evaluation metrics, i.e. R^2, RMSE, and MAE. Variable importance was extracted from the final best model to identify the predictors contributing most to variation in emissions per capita.

3. Results

3.1 Data, EDA, and Regression Analysis Summary

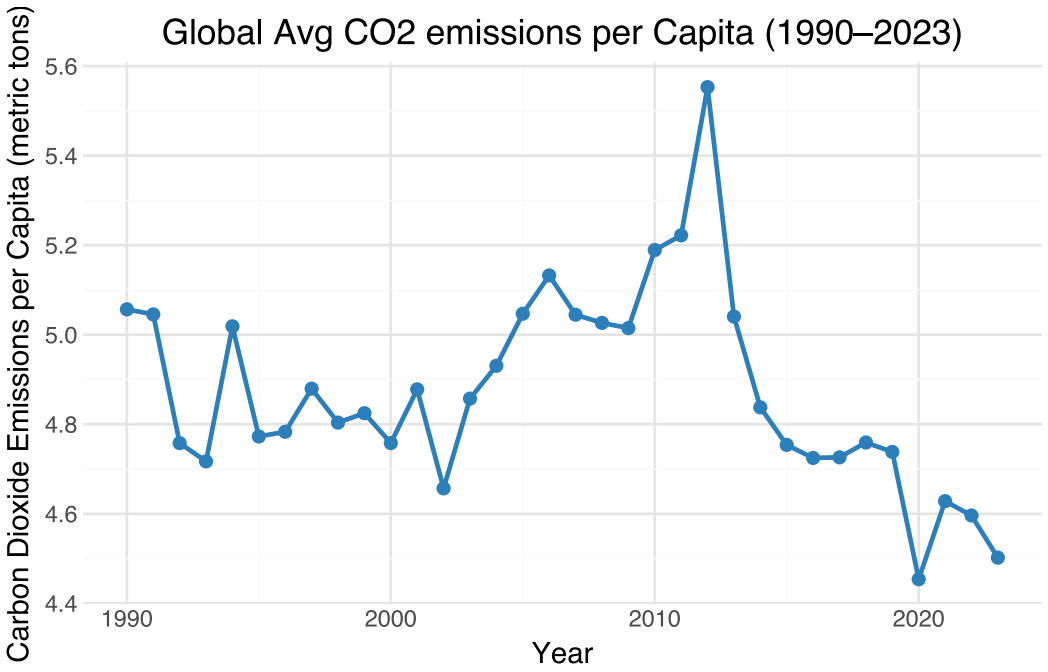

Figure 1: Global average Carbon Dioxide emissions per capita over time (1990–2023)

Table 1: GAM smooth-term summary

	Term	EDoF	p-value
0	Log GDP per Capita	16.4797	0.0
1	Fossil Electricity Share (% of total electrici...	16.0733	0.0
2	Renewable Energy Share (% of total final energ...	15.8542	0.0
3	Forest Area (% of total land area)	15.4707	0.0
4	Urban Population Share (% of total population)	15.1863	0.0
5	Year (centered at 1990)	0.9779	0.0
6	Intercept	0.0000	0.0

The cleaned dataset covers 1990–2023 after regional aggregates were removed and variable ranges were verified. Missing values were moderate (most notably for fossil-fuel electricity share) and were handled by removing all the rows with missing entries before regression analysis and modelling. The dataset had 7378 rows before removal and 5499 rows after.

According to Figure 1, carbon dioxide emissions per capita rose through the 2000s and declined after the early 2010s, with developed countries trending down and developing countries trending up in the long term. Emissions and GDP per capita were strongly right-skewed, which motivated the log transformations. According to Table 1, the GAM showed clear nonlinear relationships between emissions and the predictors (with extremely high EDoF values except for the year variable), which motivated the use of tree-based machine learning models to capture more complex patterns.

3.2 Modeling

Table 2: Country-grouped train / validation / test split

	Set	Rows	Countries
0	Train	3826	136
1	Validation	833	30
2	Test	840	30

According to Table 2, the country-level data split placed roughly 70% of countries (136 out of 196) in the training set, 15% (30 out of 196) in the validation set, and 15% (30 out of 196) in the test set, with no country has observations appearing in more than one set. Therefore, test error measures generalization to unseen countries given their full 1990–2023 predictor history, which directly addresses the cross-country focus of the research question.

Table 3: Test-set performance comparison across models

	Model	Test R^2	Test RMSE	Test MAE
0	Decision Tree (pruned)	0.7056	0.5581	0.3744
1	XGBoost (tuned)	0.7217	0.5426	0.3592

The pruned Decision Tree, after tuning `ccp_alpha` on the validation set and refitting on combined train and validation data, contained 18 leaves at depth 7 and (according to Table 3) reached a test-set R^2 of 0.706, RMSE of 0.558, and MAE of 0.374 on log carbon dioxide emissions per capita.

Table 4: Tuned XGBoost hyperparameters (5-fold GroupKFold random search + CV early stopping)

	Hyperparameter	Value
0	max_depth	6.0000
1	learning_rate	0.0300
2	n_estimators	154.0000
3	reg_lambda	0.1000
4	reg_alpha	10.0000
5	subsample	0.8000
6	colsample_bytree	1.0000
7	min_child_weight	5.0000
8	gamma	1.0000
9	CV RMSE (random search)	0.5979
10	CV RMSE (early stop)	0.5282

The tuned XGBoost used the hyperparameters in Table 4, where the hyperparameters were selected by 5-fold country-level randomized search on the training set (best CV RMSE = 0.5979) and refined by cross-validated early stopping on the combined train and validation sets, which selected 154 boosting rounds (CV RMSE = 0.5282). Refit on the combined sets, it reached a R^2 of 0.722, RMSE of 0.543, and MAE of 0.359 as the test performance. The untuned baseline XGBoost (200 trees, depth 5, learning rate 0.1) reached $R^2 = 0.681$, RMSE = 0.581, MAE = 0.388, confirming that tuned XGBoost outperformed the untuned baseline.

On the same test set (Table 3), the tuned XGBoost outperformed the pruned Decision Tree on every metric: R^2 rose from 0.706 to 0.722, and RMSE and MAE both fell from 0.558 to 0.543 log units and from 0.374 to 0.359 log units, respectively. The bettern performance across all three metrics indicates the tuned XGBoost Regressor captures more complex nonlinearities and predictor interactions that the single pruned tree could not. Therefore, the tuned XGBoost Regressor is the final best model.

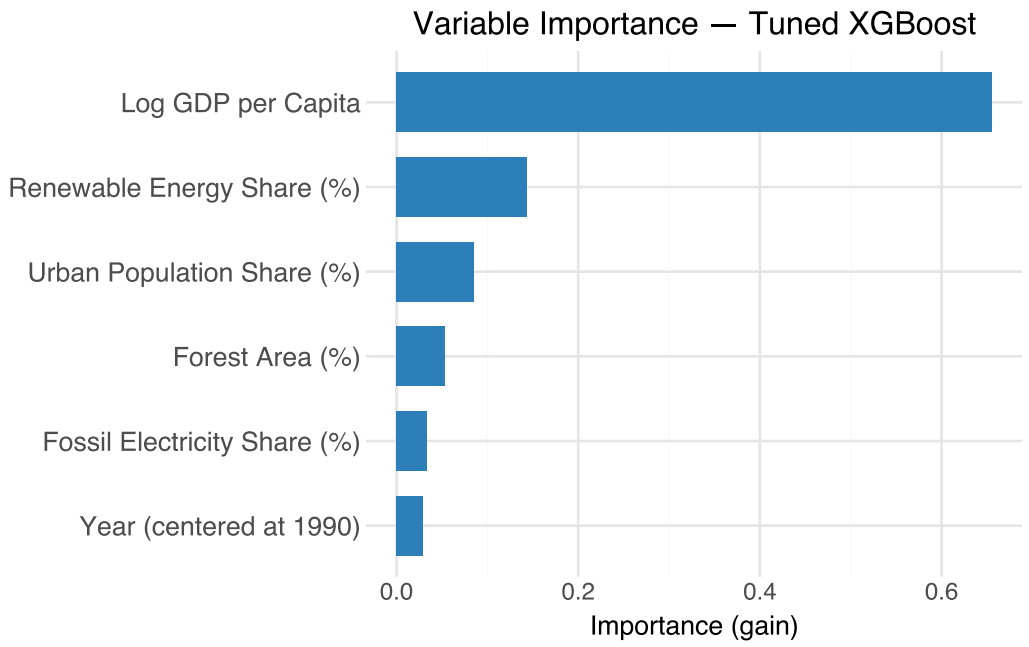

Figure 2: Gain-based feature importance for the tuned XGBoost (final model)

According to Figure 2, gain-based variable importance from the final model ranks log GDP per capita first by a clear margin, so log GDP per capita is the most important feature in determining the variation in carbon dioxide emissions per capita. Then renewable energy share and urban population share are the second and third most important features, followed by forest area, fossil electricity share, and the year. This ordering agrees with the hypothesis that economic development, energy sources, and urbanization are the key predictors of cross-country differences, while the calendar year adds little once those are known.

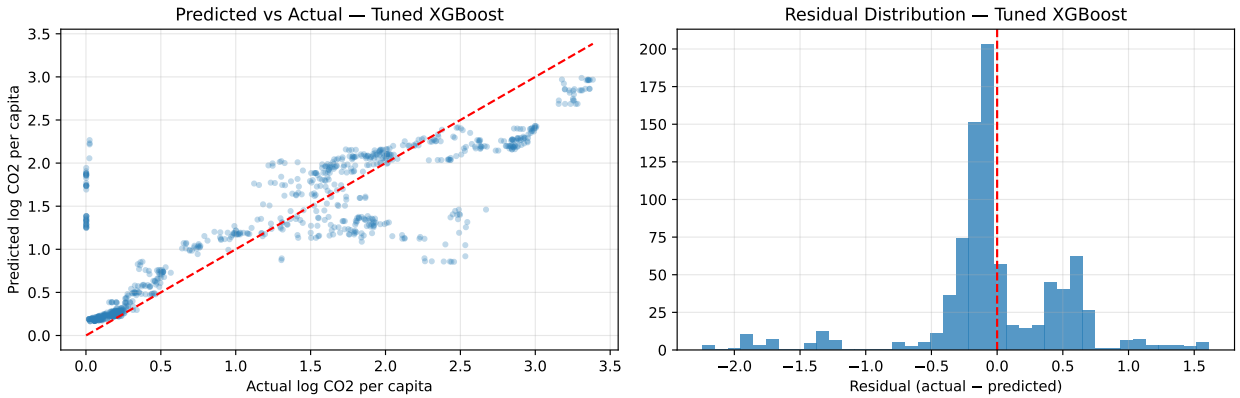

Figure 3: Predicted vs actual log carbon dioxide emissions per capita on the test set (tuned XGBoost)

According to Figure 3, the predicted outcome V.S. actual target diagnostic plot shows points clustered along the 45-degree line with no systematic curvature, and residuals are roughly symmetric and centered approximately at zero, meaning the model is not consistently too high or too low for

either low emissions or high emissions countries. The remaining spread likely reflects country related features, such as natural resources, climate changes, etc., that the available predictors do not capture directly in the dataset.

4. Conclusions and Summary

4.1 Summary of Findings and Bigger Picture

Carbon dioxide emissions per capita increased throughout the 2000s, peaked in the early 2010s, and declined afterward. The tuned XGBoost model outperformed the pruned Decision Tree and the untuned baseline on all metrics (R^2, RMSE, MAE), with residuals roughly centered around zero. Variable importance shows log GDP per capita as the dominant predictor, followed by renewable energy and urban population share, while other variables contribute much less. These results answer the research question and supports the hypothesis that economic development, energy structure, and urbanization drive cross-country differences.

Economic development remains the strongest predictor of carbon dioxide emissions per capita even after energy and land-use indicators are included. For developing countries, this highlights a key challenge: balancing economic growth with sustainability. The results suggest that this balance is more likely to come from structural changes, such as adopting cleaner energy, improving efficiency, and shifting toward low-carbon industries, rather than simply limiting growth.

Renewable energy share was the second most important predictor, while fossil-fuel electricity share contributed far less. One possible reason is that renewable share captures broader aspects of the overall energy system, not just electricity. From a policy perspective, this suggests that increasing investment in renewable energy may be one of the most effective ways to reduce emissions among the factors considered here.

4.2 Limitations

The country-level split tests generalization to unseen countries using full 1990–2023 data, not time-based forecasting. Predicting future values would require a different setup (e.g., train on 1990–2022 and test on 2023).

Dropping rows with missing entries before modelling removed a non-trivial share of country-years. If missing data is not evenly distributed across countries, this could bias the sample toward countries with better data availability.

The set of predictors is small and only reflects country-level aggregates. Some potentially relevant variables, such as energy prices or weather conditions, are not included, but they may explain part of the remaining variation in emissions. In addition, emissions are measured on a production basis, so countries that rely heavily on trade may appear differently if emissions were instead measured based on consumption.

Finally, it is important to note that the results are purely associational. Tree-based ML models do not provide causal conclusions, and the variable importance measures only indicate how useful each predictor is for making predictions within this model, not how effective a policy based on that variable would be.